COHERENCE AND PREDICTION*

by

David A. Lane

University of Minnesota

Technical Report No. 392

August, 1981

*To be presented at ISI meeting, Buenos Aires, December, 1981.

COHERENCE AND PREDICTION

DAVID A. LANE

School of Statistics
University of Minnesota
Minneapolis, Minnesota
United States of America

INTRODUCTION

This talk presents some theorems about coherence, rules for consistent reasoning in the face of uncertainty. The particular rules incorporated in these theorems, and the criteria for consistency which they enforce, arise from de Finetti's solution to the most basic problem of inference: how to measure uncertainty. de Finetti proposes a measure which reduces the assessment of uncertainty to an economic decision. Setting aside technical difficulties involving the utility of money, his measure can be defined as follows: if you want to express your degree of belief in a proposition, you decide upon a number p such that you are neutral between buying and selling for $p a ticket which is worth $1 if the proposition is true, nothing otherwise. When you simultaneously assess the uncertainty of many propositions which are related in various ways, you have set the price of many such tickets. Is it possible, in principle, that someone could transact with you for some of these, at your prices, in such a way that you must pay out more than you receive from him, no matter which of the propositions are true and which false? If so, in your assessments you have in effect made economic decisions with untoward economic consequence--sure loss--which you should be unwilling to accept. Coherence theorems delineate the constraints on your assessments which you must put into effect, if you are to avoid facing such economic catastrophe--and the inconsistent reasoning which it objectifies.

de Finetti's definition is operational, of course, only if at some point you will know whether the proposition is true or false. Hence the inferential problem to which his uncertainty measure is directly applicable is the prediction of future observables. Consequently, the theory based upon this measure makes prediction the central inferential paradigm, displacing the more common--if less securely founded--emphasis on the estimation and testing of unknown parameters.

In this talk, the de Finetti measure is adopted, at least as a model for thinking about uncertainty. Then, the coherence theorems appropriate to a variety of predictive situations, along with the implications for economic disaster which they are designed to avoid, are discussed. Section 1 presents the basic prediction result, due to de Finetti. To avoid sure loss, you must assess your uncertainty about a future observation by selecting a finitely additive probability measure on the set of possible

observed values. This result injects the mathematical theory of measure into the theory of inference; in contrast, Jeffrey's development starts with the unargued assumption that uncertainty measures must conform to the usual axioms of probability theory. Of course, the mathematical theory invoked by de Finetti's theorem is about finitely- but not necessarily countably-additive measures, which introduces complications for those accustomed to Kolmogorov's axiomatization.

In Section 2, the operational character of de Finetti's uncertainty measure is abandoned. Instead, the measure is used to motivate a model for the statistical problem of inference about an unknown parameter. The main result--due to Heath and Sudderth--is that such inference is coherent if and only if it is consistent with the posterior calculated from some finitely additive prior distribution on the parameter space. The coherence criterion used in this setting is weaker than sure loss and depends upon public agreement that the observation is generated from some distribution in the specified parametric family. Basically, this criterion calls for the avoidance of a uniformly unfavorable economic horizon, considered over all the specified possible mechanisms for generating the observation.

Section 3 returns to the problem of predictive inference: you are going to see two observations in sequence, and you want to know how knowledge of the first should shape your prediction of the second. Two theorems are presented. The first relates predictions made with and without knowledge of the first observation, and requires that you use a form of the law of total probability to avoid sure loss. The second assumes that a set Θ of possible joint distributions for the two observations has been publicly specified, and concludes that, to avoid uniformly unfavorable horizons before either observation is taken, you must agree to predict the second, after having seen the first, in a Bayesian manner, involving the selection of a prior probability distribution on Θ.

Section 4 is a mathematical aside, showing how to prove coherence theorems, and Section 5 contains an application of some of the theorems in the context of sampling from a finite population.

The theorems are not stated in their full generality, or with attention to the strictest mathematical detail--see the original papers for that. But one important generalization should be pointed out: it is not necessary to require that <u>every</u> event have its uncertainty measured, as is done here--the theorems remain true, but call for the existence of some measure which agrees with the partial set of assessments made. For details, see de Finetti (1972) and Lane and Sudderth (1981a).

1. COHERENT PREDICTION

You are to make an observation, and you wish first to predict what you will observe. How should you express your prediction?

First, you must specify the possible observations: a set X. Next, you should distinguish among these possibilities, according to how likely you think each is to occur. You can do this, following de Finetti, in this way: for each subset S of X, determine a number, $p(S)$, such that you would be neutral between buying and selling a ticket which costs $\$p(S)$, and is worth $\$1$ if S happens, nothing otherwise.

Some prediction functions have self-contradictory implications. For example, suppose a weatherman specifies X = {rain tomorrow, no rain tomorrow}, and he assesses $p(\text{rain tomorrow}) = p(\text{no rain tomorrow}) = 1/3$. If I

were to take him up on his evaluations, and buy from him a ticket on rain and a ticket on no rain, for an outlay \$2/3 I would have a pair of tickets which together would be worth \$1 tomorrow, whether it rained or not. This kind of inconsistency is called incoherence.

More generally, suppose X is any set and p a prediction function on X. Make p the basis for a book, and allow a gambler to place a finite number of bets from this book. That is, the gambler may select n subsets of X--say $S_1,\ldots,S_n$--and n real numbers--$\alpha_1,\ldots,\alpha_n$. He lays out $\$\Sigma_{i=1}^{n}\alpha_i \cdot p(S_i)$, and then, for each i, he collects $\$\alpha_i$ if S_i occurs, nothing otherwise. (A negative α reverses the roles of buyer and seller between gambler and predictor.) If, for some $c > 0$, the gambler can select sets and stakes in such a way that he wins at least $\$c$, no matter what is observed, then the predictor and his prediction function p are called incoherent. Otherwise, they are coherent. For a justification of this formulation--in particular, of the finiteness of n and the strict positivity of c--see de Finetti (1974).

According to a fundamental result of de Finetti, a prediction function p is coherent if and only if it satisfies:

i) for each subset S of X, $0 \leq p(S) \leq 1$,

ii) for each pair of disjoint subsets S_1 and S_2 of X,

$$p(S_1 \cup S_2) = p(S_1) + p(S_2).$$

According to this result, to solve the problem of prediction, you select a finitely additive probability distribution for the future observation. This distribution summarizes your opinion--based upon the information available to you--about what you expect to observe. You have a lot of choice: there are many finitely additive probability measures on X. But should you frame any predictive statement which is not consistent with at least one of these, you in effect open the possibility of a Dutch book against yourself.

2. STATISTICAL COHERENCE

In the standard statistical set-up, in addition to the set X of possible observations, the observer specifies a set Θ. Each θ in Θ corresponds to a probability distribution, p_θ, on X. The interpretation is that θ specifies a state of information about the process generating the forthcoming observation. If the observer possesses the information specified by θ, he would use p_θ as his prediction function for the observation. Instead, the information available to the observer leads him to believe that one element of Θ actually describes the process generating the observation, but he is uncertain about which element this is. The problem of inference in this situation is retrodictive: after observing x, what should the observer conclude about which θ in Θ actually describes the process which generated x?

Superficially, this problem resembles the question considered in the previous section. Why not determine, for each subset S of Θ, a price $\$p(S)$ for a ticket worth \$1 if the correct θ is in S, nothing otherwise--and then decide what properties such a p must have to be coherent? Unfortunately, such a program cannot be realized operationally, even conceptually: θ represents a state of information, not a future observable,

and so bets on subsets of Θ can never be settled up. This is why de Finetti describes the standard statistical formulation as metaphysical; it is also at the heart of Geisser's propaganda for predictivism as the preferred mode of statistical analysis.

Nonetheless, Freedman and Purves (1969) have suggested a model for statistical inference which involves acting as though θ were observable. For their model, Freedman and Purves invent a scenario with three characters: a master of ceremonies (MC), a bookie (our observer--or inferrer, as he should now be regarded), and a gambler. All three agree upon X and Θ. The MC arbitrarily selects θ in Θ and then, using the distribution p_θ, he generates an observation in X. Before the MC reveals the value of his observation, the bookie announces a book on subsets of Θ for each x (based upon a prediction function q_x), and the gambler, after studying these books, decides upon a finite number of bets from each of them. The MC then declares the value he observed, say x: so now only one of the books, the one based upon q_x, and a finite number of bets, those based upon this book, are relevant. The gambler places these bets, the MC reveals which θ he used to generate x, and the gambler and bookie settle up.

The set of predictive functions $\{q_x : x \text{ in } X\}$ are the important objects in this scenario: q_x represents the inferrer's opinions about Θ consequent upon the observation of x. What constraints should these functions satisfy? First, the inferrer should not offer a Dutch book, no matter what x is observed. By de Finetti's result, this will be satisfied if and only if each q_x is a finitely additive probability measure on Θ. This condition takes no account of the assumed relationship between the observation and the set of possible generating mechanisms represented by Θ.

To see that something more is necessary, consider this example. Both Θ and X are the integers, and $p_\theta = \frac{1}{2}(\delta_{\theta-1} + \delta_{\theta+1})$, where δ_y is point mass at y. If x is observed, θ must be either $x-1$ or $x+1$. As long as $q_x(\{x-1\})$ and $q_x(\{x+1\})$ add to unity for some x, the gambler cannot win for sure. Suppose the bookie decides to make $q_x = \frac{1}{3}\delta_{x-1} + \frac{2}{3}\delta_{x+1}$. The gambler responds by betting on $\{(\theta,x): x = \theta+1\}$; that is, if x is observed, he will pay $\$\frac{1}{3}$ for a ticket on $\{x-1\}$ and will receive $\$1$ back if θ is $x-1$, nothing otherwise. Now with this betting system, the gambler achieves an advantage over the bookie: all participants know that some θ has been chosen; for anyone who knows this θ, the expected gain for the gambler is $\frac{1}{2}(\$\frac{2}{3}) + \frac{1}{2}(-\$\frac{1}{3}) = \$\frac{1}{6}$; but this expected gain does not depend on θ. Thus, if the bookie assesses his economic situation before x is revealed, from the point of view of any of the possible values for θ, his horizon is cloudy; he foresees a loss of $\$\frac{1}{6}$. No matter what rule governs the game, to the bookie, the game is unfavorable!

In general, call the bookie and his set of prediction functions $\{q_x : x \text{ in } X\}$ <u>coherent</u> if:

1) Each q_x is coherent in the sense of section 1; and

2) there is no betting system available to the gambler, under which the expected loss of the bookie, calculated as though θ were known to have been selected, is greater than \$c, for every θ and some strictly positive c.

Otherwise, the bookie and his prediction functions are <u>incoherent</u>. To the extent that this scenerio is a convincing model for statistical inference, an inferrer should behave as a coherent bookie. Refer to the $\{q_x : x \text{ in } X\}$ of the inferrer as his set of inferential distributions, and call this set coherent if the corresponding Freedman-Purves prediction functions are coherent.

Heath and Sudderth (1978) have characterized coherent inferential distributions. Suppose each q_x is a finitely additive probability measure on Θ. In addition, suppose π is a finitely additive probability measure on Θ, and for all bounded real-valued functions f on $\Theta \times X$,

$$\int\int f(\theta,x)p_\theta(dx)\pi(d\theta) = \int\int f(\theta,x)q_x(d\theta)m(dx)$$

for some finitely additive probability measure m on X. Then $\{q_x : x \text{ in } X\}$ is called a <u>posterior</u> for the <u>prior</u> π.

<u>Theorem</u> (Heath-Sudderth): A set of inferential distributions $\{q_x : x \text{ in } X\}$ is coherent if and only if it is a posterior for some prior π.

According to this theorem, coherent inference is Bayesian inference: select a prior, observe x, and compute the posterior on Θ given x. If Θ and X are both finite, Bayes' formula for computing posteriors by multiplying likelihood and prior is always available. So the problem of coherent inference is completely solved by selecting a prior measure on Θ: the rest is mechanical application of Bayes' formula. If θ is infinite, it supports finitely additive measures which are not countably additive, and complications arise. In particular, there are finitely additive priors and countably additive likelihoods which yield no posteriors at all. Moreover, there is not necessarily a simple algorithm like Bayes' formula for computing the posterior from a finitely additive prior, even if it does exist. There are in general many posteriors for a given prior. Despite all this, the theorem is a useful one. If, after all, the inferrer wishes to be coherent, the theorem gives him a necessary and sufficient condition for being so, and he only needs to work out the mathematics to verify that his favorite set of inferential distributions can in fact be obtained as the posterior for some finitely additive prior. Here is an example which has many statistical applications. Suppose Θ and X are the same amenable group, and the sampling distributions p_θ form a translation family (see Heath and Sudderth for definitions; as a simple example to keep in mind, let $\Theta = X = R$ and p_θ be $N(\theta,1)$). Then the formal Bayes posterior with respect to left Haar measure as prior is coherent. But if any other relatively invariant prior is used, the resulting invariant posterior is incoherent (see Lane and Sudderth (1981c)).

In standard countably additive Bayesian analysis on infinite parameter spaces, improper priors are frequently employed, and Bayes' formula is

formally applied to produce a posterior. Improper priors do not meet the specifications of the Heath-Sudderth theorem, so the question arises: are the posteriors obtained by this recipe coherent? Sometimes yes, sometimes no, as the result mentioned in the last paragraph implies. Some general results classifying which improper priors lead to coherent formal posteriors are discussed in Lane and Sudderth (1981b). Another question involving improper priors is: do all coherent inferential distributions arise as formal posteriors from some--perhaps improper--countably additive prior? The answer is no; some examples arise in connection with the marginalization paradox (Dawid, Stone, and Zidek (1973)) - see Sudderth (1980).

3. COHERENT PREDICTIVE INFERENCE

Consider taking successive observations which are related to each other in some way. For example, you might first choose a sample of size ten from some population and then select an additional item, or you might measure two different physical constants whose actual values are connected by some theory. Call the sets of possible values for the first and second observations X and Y respectively. For convenience--and hopefully without confusion--also refer to the observations themselves, before they have been taken, as X and Y.

How should your predictions about Y be modified by what you actually see in the first observation? This is the problem of predictive inference. We shall consider two aspects of this problem. First, what relations should obtain between your opinions about Y before and after the first observation is taken? The second aspect involves the actual mechanism of opinion modification. Suppose a set Θ has been specified, whose elements characterize the process which jointly generates the two observations. Seeing x on the first observation gives information about which θ in Θ provides the correct characterization, and this information is perhaps different for different possible values of X. How should you convert the information about Θ you gain from the first observation into a predictive distribution for the second?

To answer the first question, suppose you record all your thoughts about the two observations before either is taken. You begin by specifying X and Y, and you determine the following prediction functions, to be interpreted as described in Section 1: p_X and p_Y, on X and Y respectively, which express your current opinions about the two observations; and for each x in X, the prediction function q_x on Y, which expresses what your opinions about the second observation will be, should it turn out that x is the value of the first one.

Now, to determine whether these prediction functions are mutually consistent, you are required to convert them into bets and face the economic consequences of your opinions. What constitutes an allowable betting system for the gambler who uses your books? He should, as always, be restricted to a finite number of cash transactions, and he should announce all his bets before either observation is taken. The finite number of bets from the books based upon p_X and p_Y should be paid for before either observation is taken, since they represent opinions held at that time. What about bets from the books based upon the prediction functions $\{q_x : x \text{ in } X\}$, representing your opinions after the first observation?

Since the bets based upon q_x will be in force only if x is observed, it is most reasonable to have the cash transactions for these bets occur after the first observation (if it is x!) but of course before the second. This is premised upon the sequential character of the observations, which is central in the predictive inference framework. (Contrast this with the set-up used by de Finetti when he proves that, if A and B are events depending on the same observation, then $p(A \cap B) = p(B|A)p(A)$--here, of course, the cash transactions for all three relevant bets are concluded at the same time). Thus, if the gambler has chosen to stake $\$\alpha$ on a subset S of Y using the book derived from q_x, he pays you $\$\alpha \cdot q_x(S)$ if and when x is observed, and he collects $\$\alpha$ only if the second observation is in S. If x is not observed, there is no transaction. Thus, while the gambler has announced as part of his betting system a finite number of bets from each q_x - book, he is always restricted to a finite number of cash transactions (and collectible bets).

With these rules, the predictor and his prediction functions are coherent if

1) there is no system of bets available to the gambler which guarantees him a win of $\$c$ $(c > 0)$, no matter which x and y are observed; and
2) for each possible value x of the first observation, there is no system of bets on Y available to the gambler which guarantees him a win of $\$c$ $(c > 0)$, no matter which y is observed.

So if the predictor is coherent, neither before or after the first observation will he face a situation in which he owes the gambler a certain sum no matter what the future holds.

By de Finetti's result, all of the predictions--p_X, p_Y, and each q_x--must be finitely additive probability measures on their respective domains. They must also fit together as specified by the following theorem due to Lane and Sudderth (1981a).

Theorem: The set of prediction functions described above is coherent if and only if:

1) p_X is a finitely additive probability measure on X,
2) for each x in X, q_x is a finitely additive probability measure on Y, and
3) for each subset S of Y,

$$p_Y(S) = \int q_x(S) p_X(dx).$$

Condition 3) shows that in the predictive framework considered here, the law of total probability, like finite additivity, is derivable from the criterion of avoiding sure loss. Had the gambler been required to pay for all potential transactions (from each q_x - book) before the first observation, condition 3) would be necessary only in case the set X is finite.

Not all subjectivists will agree with the formulation adopted here, as the following example will perhaps indicate. This example relates to the famous Kolmogorov-Borel paradox involving surface area on spheres, which has been discussed from the subjectivist viewpoint by de Finetti

(1974) and Hill (1979). Here is a statement of the example in the predictive inferential framework. Someone selects a point on the unit sphere. You will first observe its longitude relative to a given polar axis, so X is $[-\pi,\pi]$. Next, you will observe its exact position, so Y is the surface of the sphere. You feel that the point has been selected completely "at random", so you evaluate p_X and p_Y as uniform on (the Borels of) X and Y respectively. Given the longitude of a point, its position is determined by its latitude, so you let q_x be the distribution on Y concentrated on points with longitude x and latitude uniformly distributed on $[-\pi/2,\pi/2]$. (If you do not find these assessments compelling, Hill presents reasons why perhaps you should.)

Unfortunately, these prediction functions do not satisfy condition 3). This has nothing to do with finite additivity, since all the distributions involved are countably additive (on Borels). Rather, the problem hinges on what you mean by the assessments conditional on longitude. I contend that this should depend on how you intend taking the observations and what you want to make inferences about, since the method whereby uncertainty is assessed is designed to provide an operational interpretation of a particular inferential situation. (For example, there is no inconsistency in the weatherman introduced in section 1 assessing $p(\text{rain}) = 1/3$, unless he also assesses $p(\text{no rain}) = 1/3$.) If you envision taking observations and making inferences about them within the sequential scheme outlined here, you should regard the "all uniform" assessment, whatever merits some of its components might have as assessments in other settings, as jointly unacceptable in this one.

For another aspect of the relation between coherence and the particular inferential framework in which you make your assessments, suppose you choose to regard X and Y as a single observation (X,Y). You can coherently predict (X,Y) by specifying any finitely additive probability on $X \times Y$. Suppose you choose μ. Now you have not been called upon to predict Y after X has been observed; this additional task would require additional selections, namely the set of prediction functions $\{q_x : x \text{ in } X\}$. It would be nice if your prediction for (X,Y) necessarily also equipped you for the coherent sequential prediction of the two observations. This would be the case if μ disintegrates with respect to the product structure on $X \times Y$: that is, if there exist finitely additive measures $\{q_x : x \text{ in } X\}$ on Y such that, for each subset A of $X \times Y$,

$$\mu(A) = \int q_x(A_x) p_X(dx),$$

where $A_x = \{y: (x,y) \text{ is in } A\}$ and p_X is the marginal induced by μ on X. Unfortunately, this need not be the case: see Dubins (1975). On the other hand, if you intend making sequential predictions, and have chosen coherent prediction functions p_X, p_Y and $\{q_x : x \text{ in } X\}$, then you have available to you a coherent prediction for the combined observation (X,Y): namely, the strategic measure μ on $X \times Y$ determined by the integral displayed above.

We now turn to the second problem mentioned above. A set Θ is specified, whose elements characterize the process which jointly generate

X and Y. You believe one of these elements correctly characterizes this process, but you are uncertain as to which one it is. If you could determine that θ were correct, you would use p_θ as your joint predictive distribution for the two observations. The problem is to use information about θ which you have after observing x to decide upon your prediction for Y.

For this problem, only your predictions after the first observation are of interest. So you begin by announcing a predictive distribution on Y, q_x, for each possible value x of the first observation. These distributions are converted into books, and the gambler announces a finite number of bets from each of them. After the first observation is taken, the relevant bets are put into effect; the second observation is then taken, and accounts are settled between you and the gambler.

Before X or Y is observed you can contemplate the gambler's announced betting system and calculate--for any θ in Θ--what you would expect to lose if θ characterized the generating process correctly. Suppose you calcualted this number for <u>each</u> θ, and you always obtained a figure exceeding $\$c$ $(c > 0)$. Then, no matter which characterization were accurate, the gambler appears from the point of view of the correct θ to hold a fixed advantage over you. If you wish to avoid unfair games of this sort, you will make their impossibility a criterion for a coherent assignment of predictive distributions. Let us do so.

Then, the predictor and his prediction functions are, by definition, <u>coherent</u> if:

1) for each x in X, there is no system of bets on Y available to the gambler which guarantees him at least $\$c$ $(c > 0)$, no matter which y is observed.
2) There is no system of bets on Y available to the gambler which gives him expected winnings greater than $\$c$ $(c > 0)$, calculated according to p_θ, for each θ in Θ.

To satisfy condition 1), you must be sure that each q_x is a finitely additive measure on Y. What about condition 2)? Based upon the Heath-Sudderth theorem, it is reasonable to expect a Bayesian solution to this problem; indeed, this is so. The condition--due to Lane and Sudderth (1981a)--is as follows:

<u>Theorem</u>: A set of prediction functions $\{q_x : x \text{ in } X\}$ is coherent if and only if

1) each q_x is a finitely additive probability measure on Y and
2) there exists some finitely additive measure π on Θ such that the measure m induced on $X \times Y$ by π and $\{p_\theta : \theta \text{ in } \Theta\}$ is strategic, with conditionals $\{q_x : x \text{ in } X\}$. That is, for any subset A of $X \times Y$,

$$m(A) = \int p_\theta(A)\pi(d\theta)$$

$$= \int q_x(A_x)m_X(dx),$$

where $A_x = \{y: (x,y) \text{ is in } A\}$ and m_X is the marginal distribution on X induced by m.

It is interesting to note that there need not be a posterior distribution on Θ given x in order for condition 2) to hold. That is, there are situations in which you may choose a prior which allows you to predict the next observation coherently, but does not allow you to "estimate" the value of the parameter coherently! An example can be constructed based upon the models for sampling from a finite population introduced in Lane and Sudderth (1978). For such an example, Θ would be the set of all real-valued finite populations (i.e., $U_{n=1}^{\infty} R^n$). The first observation would be, say, the values of four items sampled from one of these populations (so $X = R^4$). Suppose you then put the values of these four items in order, creating five intervals on the real line (from $-\infty$ to the smallest value, the smallest value to the next smallest, etc.). Take a fifth item from the population. Your second observation tells you which of those five intervals this item is in (so $Y = \{1, 2, 3, 4, 5\}$, where 1 specifies the smallest interval, etc.). Each θ in Θ specifies the population from which the sample is selected, and p_θ says that the sample is selected at random without replacement from the population θ. Now there exist prior distributions on Θ, such that the following assignments would give coherent prediction functions: no matter which x is observed (i.e., which four items are selected),

$$q_x\{1\} = q_x\{2\} = \ldots = q_x\{5\} = 1/5.$$

There are no posteriors on Θ for these distributions, given the values of samples of size four selected from the population; nor could you even use them to predict coherently the value of the fifth item sampled. This example will be amplified in section 5.

A final note about this coherence theorem: the predictions involved are about observables (Y), based only on observed values (x in X). Yet whether the predictions are coherent depends upon the specification of the set Θ. As an example, suppose two successive measurements of the same quantity, using the same apparatus, are contemplated. First, suppose measurement error is disregarded, so Θ contains only distributions concentrated on the diagonal of R^2. Then, unless q_x is δ_x, the predictions must be incoherent. Clearly, if Θ specifies instead independent measurements from some nondegenerate translation family, there is a much larger class of coherent prediction functions.

4. WHY ARE COHERENCE THEOREMS TRUE?

Suppose A and B are disjoint convex sets in the Euclidean plane, and A has nonempty interior. Then you can always separate A and B by a line. This fact generalizes: according to the so-called separating hyperplane theorem, it is true if the sets are in a general locally convex topological vector space. In this setting, the theorem asserts that dis-

joint convex sets A and B, at least one with nonempty interior, can be separated by a hyperplane. Equivalently, there exists a nonzero continuous linear functional π on the space and a real number r, such that the value of π is at least r for each element of A, and at most r for each element of B. The separating hyperplane theorem is an easy consequence of the Hahn-Banach theorem (see Reed and Simon (1972), p. 130), and it, in turn, can be used to prove each of the coherence theorems discussed above.

To use the separating hyperplane theorem, it is helpful to rephrase its conclusion in a special case. Suppose X is a set. Let $L(X)$ consist of the bounded real-valued functions on X, and equip $L(X)$ with the sup norm. Let A be the uniformly positive functions in $L(X)$:

$$A = \{f \text{ in } L(X): \sup_{x \text{ in } X} f(x) > 0\}.$$

Also, let B be a subspace of $L(X)$ (not necessarily closed). Now suppose A and B are disjoint. Since they are both convex, and A has nonempty interior (for example, it includes the function with constant value 1), the separating hyperplane theorem implies that there is a continuous linear functional which is at least r on A and at most r on B. Moreover, since the constant function whose value is 0 is a limit point of both A and B, the number r must be 0. Then, by continuity, the functional must assign a nonnegative value to every nonnegative function. Each nonnegative linear functional on $L(X)$ corresponds to integration with respect to a unique nonzero bounded finitely additive measure on X. Let π denote the measure whose integral separates A and B. Since B is symmetric ($B = -B$), for every f in B, it must be the case that $\int f \, d\pi = 0$. The following lemma sums up the results of this paragraph:

Lemma: X is a set, $L(X)$ the space of bounded functions on X equipped with the sup norm, A the cone of uniformly positive functions in $L(X)$, B a subspace of $L(X)$. If A and B are disjoint, there is a nonzero bounded finitely additive measure π on X which integrates every function in B to zero.

To illustrate how this lemma applies to coherence, here is a proof of the theorem on coherent prediction discussed in Section 1. X is the set of possible observations. B consists of all possible net winnings for a gambler using the book based upon the prediction function p. Thus, B is the finite linear span of functions of the form

$$g_S(x) = 1_S(x) - p(S)$$

where S is a subset of X. By definition, p is coherent if and only if no function in B is uniformly positive: that is, if B is disjoint from the cone of uniformly positive functions in $L(X)$. Thus, if p is coherent, the above lemma can be applied. This yields a nonzero bounded finitely additive measure π on X which integrates each function in B to zero: in particular, for each subset S of X,

$$0 = \int g_S d\pi$$

$$= \pi(S) - p(S)\pi(X).$$

That is,

$$p(S) = \frac{\pi(S)}{\pi(X)} .$$

So p is a finitely additive probability measure on X. The converse result may be obtained directly: if p is not additive, the bookie can be exploited just as in the weatherman example of Section 1.

The first coherence theorem of Section 3 is proved similarly to the above. For the second theorem of that section and the Heath-Sudderth result of Section 2, the relevant coherence condition involves expected winnings given θ, and so the relevant function space is $L(\Theta)$. For example, here is a sketch of the proof of the Heath-Sudderth result. Suppose the gambler chooses a single set in Θ to bet on for each possible x: call the chosen set A_x. Now define the set A on $\Theta \times X$ by

$$A = \{(\theta,x): \theta \text{ is in } A_x\}.$$

(Of course, any A in $\Theta \times X$ can be obtained in this fashion). With this betting system, the gambler's winnings W are

$$W(\theta,x) = 1_A(\theta,x) - q_x(A_x).$$

His expected winnings f, given θ, are

$$f(\theta) = \int 1_A(\theta,x) p_\theta(dx) - \int q_x(A_x) p_\theta(dx).$$

Let B be the finite linear span of such functions f on Θ. The coherence criterion implies that, if the assignment $\{q_x: x \text{ in } X\}$ is coherent, B is disjoint from the cone of uniformly positive functions.

Then, assuming coherence, apply the lemma to get a finitely additive probability measure π on Θ whose integral annihilates B. In particular, for each A in $\Theta \times X$,

$$\iint 1_A(\theta,x) p_\theta(dx)\pi(d\theta) = \iiint 1_A(\theta,x) q_x(d\theta) p_\theta(dx)\pi(d\theta)$$

$$= \iint 1_A(\theta,x) q_x(d\theta) m(dx)$$

where m is the marginal on X induced by π and $\{p_\theta: \theta \text{ in } \Theta\}$. Now the displayed equality can be extended from indicators to all bounded functions on $\Theta \times X$, and it then would show that $\{q_x: x \text{ in } X\}$ is indeed the posterior for the prior π. The converse is again easy.

5. DIFFUSE OPINION AND COHERENT INFERENCE

In each of the situations described in Sections 1 through 3, coherence required the selection of finitely additive probability measures, either as prediction functions on an observation space or as prior distributions on a parameter space. Most statisticians are familiar with quite a few countably additive distributions, and if opinions are so sharp that they can be described by one of these, all is well. However, some opinions are too diffuse to be represented by any countably additive distribution. In such cases, probability distributions which are not countably additive are needed. Unfortunately, few of these are well-known to statisticians. So an important task facing the statistician who wants to be coherent is to assemble a catalogue of finitely additive distributions which give mathematical expression to the kind of diffuse opinions which arise in common inferential situations.

This section describes a class of distributions which can be applied to some problems connected with sampling from a finite population. These distributions may be used as prior distributions on the values of some numerical characteristic which will be observed for the individuals comprising a sample selected from the population. The state of information which these distributions are designed to express about the population, the numerical characteristic, and the way in which the sample is to be chosen, has three important components, which were first isolated and described by Hill (1968). These are:

1) The possibility of precise measurement: the values of the numerical characteristic for any two individuals in the population can be distinguished. Thus, no matter how large a sample is selected, ties occur with probability zero.
2) The irrelevance of sample order: if a sample of size n is to be selected from the population, A is a subset of R^n, and π a permutation of $\{1,\ldots,n\}$, then the observed sample values are as likely to be in A as in πA.
3. Diffuse opinion about the distribution of the characteristic in the population: knowledge about this distribution is so limited that the predicted rank of the next value sampled does not depend on the values which have already been observed. That is, in a sample of size n, the probability that the n^{th} individual sampled will have a value between the i^{th} and $(i+1)^{st}$ largest of the $(n-1)$ values already observed is $1/n$, regardless of what those $(n-1)$ values were.

These three properties--especially the third--may not describe anyone's opinions exactly, yet they may provide a far more satisfying approximation than any distribution which postulates more knowledge about the relation between the scale of measurement for the characteristic and the distribution of the characteristic in the population. For example, if you are unfamiliar with physics and astronomy, you might think about your opinions regarding the population of distances between all pairs of subatomic particles in the universe, before you are to be told the values for some such distances by a capricious physicist. For a careful justification of property 3), (including the notion of a "rubbery" scale of measurement, which is not a feature of the particle distance example), see Hill (1968, p. 680).

There are many distributions which incorporate the three properties posited above. They are characterized in Lane and Sudderth (1978). Some of these distributions describe very vague ideas about the order of magnitude of the numbers to be observed: they produce samples all of whose values lie outside any given compact set with probability one. On the other hand, some of the distributions are highly concentrated: they produce samples whose values are all arbitrarily close to some fixed value with probability one. All the distributions incorporating the three properties are mixtures of these two types, as the following result shows: if a sample of size n is selected, using any of these distributions as prior on the space of populations, the unconditional probability that the range of sample values is between ε and M is zero, for any positive ε (however small) and M (however large).

From an inferential point of view, the most interesting aspect of these distributions is that they are not strategic. In particular, it is not possible to use these distributions to give coherent predictions for the value of the next individual sampled, based on a set of observed values. Nor can the values of all the unsampled units be coherently inferred. However, it is possible to make coherent predictions about the percentage of future observations falling between any two successive observed values and to make coherent inferences about some characteristics of the percentiles of the population undergoing sampling. See Hill (1968, sections 3 and 4) for details. It is interesting to note that the predictions and inferences which he derives depend on the prior distribution only through the mathematical expression of the three properties described above, and not on the particular form of the prior itself.

Finally, it should be noted that diffuse models for random sampling from a continuous cumulative distribution function can be constructed which are closely related to the diffuse distributions described above. Work in progress indicates that these distributions may be used to show that predictive distributions based upon the empirical distribution and kernel density estimates are incoherent.

6. BIBLIOGRAPHICAL NOTES

Buehler (1976) presents an interesting treatment of coherence which is not based on de Finetti's uncertainty measure. In addition, his paper records several useful separation theorems and its bibliography refers to some other important works on coherence. Stone has written many papers which relate to coherence ideas; the one most related to this talk is listed below.

BIBLIOGRAPHY

Buehler, Robert (1976), "Coherent Preferences," Annals of Statistics, Vol. 4, 1051-1064.

Dawid, A. P., M. Stone and J. V. Zidek (1973), "Marginalization paradoxes in Bayesian and structural inference," Journal of the Royal Statistical Society B, Vol. 35, 189-233.

de Finetti, Bruno (1972), Probability, Induction and Statistics (John Wiley, New York).

de Finetti, Bruno (1974), Theory of Probability (John Wiley, New York) (English translation, first published in Italian, 1970).

Dubins, Lester (1975), "Finitely additive conditional probabilities, conglomerability, and disintegrations," Annals of Probability, Vol. 3, 89-99.

Freedman, David and Roger Purves (1969), "Bayes method for bookies," Annals of Mathematical Statistics, Vol. 40, 1177-1186.

Geisser, Seymour (1971), "The inferential use of predictive distributions," in B. P. Godambe and D. A. Sprott, eds., Foundations of Statistical Inference (Holt, Rinehardt, and Winston, Toronto).

Heath, David and William Sudderth (1978), "On finitely additive priors, coherence, and extended admissibility," Annals of Statistics, Vol. 6, 333-345.

Hill, Bruce (1968), "Posterior distribution of percentiles: Bayes Theorem for sampling from a population," Journal of the American Statistical Association, Vol. 63, 677-691.

Hill, Bruce (1979), "On some statistical paradoxes and non-conglomerability," presented at the International meeting on Bayesian Statistics, Valencia, Spain.

Lane, David and William Sudderth (1978), "Diffuse models for sampling and predictive inference," Annals of Statistics, Vol. 6, 1318-1336.

Lane, David and William Sudderth (1981a), "Coherent predictive inference," University of Minnesota School of Statistics Technical Report.

Lane, David and William Sudderth (1981b), "Improper priors and coherent inference," University of Minnesota School of Statistics Technical Report.

Lane, David and William Sudderth (1981c), "Coherence and invariance," University of Minnesota School of Statistics Technical Report.

Reed, Michael and Barry Simon (1972), Functional Analysis (Academic Press, New York).

Stone, Mervyn (1976), "Strong inconsistency from uniform priors," Journal of the American Statistical Association, Vol. 71, 114-116.

Sudderth, William (1980), "Finitely additive priors, coherence, and the marginalization paradox," Journal of the Royal Statistical Society B, Vol. 42, 339-341.

SUMMARY

Opinions are coherent when they are free from internal inconsistencies. These inconsistencies may be objectified in economic terms, by demanding that opinions carry economic consequences. In this framework, criteria for coherence appropriate to a variety of predictive and inferential situations are developed, and constraints upon opinions necessary and sufficient to meet these criteria are established. The main conclusion is: coherent prediction and inference require Bayesian methodology. Diffuse opinions must be expressed in terms of purely finitely additive probability distributions. Some differences between techniques using these distributions and those based upon improper countably additive distributions are pointed out. Finally, some diffuse models for sampling from a finite population are discussed.